Proceedings of Machine Learning Research 209:37–49, 2023 Conference on Health, Inference, and Learning (CHIL) 2023

Bayesian Active Questionnaire Design for Cause-of-Death Assignment Using Verbal Autopsies

Toshiya Yoshida TOYOSHID@UCSC.EDU
University of California Santa Cruz, USA

Trinity Shuxian Fan FANSX@UW.EDU
University of Washington, USA

Tyler McCormick TYLERMC@UW.EDU
University of Washington, USA

Zhenke Wu ZHENKEWU@UMICH.EDU
University of Michigan, USA

Zehang Richard Li LIZEHANG@UCSC.EDU
University of California Santa Cruz, USA

Abstract

Only about one-third of the deaths worldwide are assigned a medically-certified cause, and understanding the causes of deaths occurring outside of medical facilities is logistically and financially challenging. Verbal autopsy (VA) is a routinely used tool to collect information on cause of death in such settings. VA is a survey-based method where a structured questionnaire is conducted to family members or caregivers of a recently deceased person, and the collected information is used to infer the cause of death. As VA becomes an increasingly routine tool for cause-of-death data collection, the lengthy questionnaire has become a major challenge to the implementation and scale-up of VA interviews as they are costly and time-consuming to conduct. In this paper, we propose a novel active questionnaire design approach that optimizes the order of the questions dynamically to achieve accurate cause-of-death assignment with the smallest number of questions. We propose a fully Bayesian strategy for adaptive question selection that is compatible with any existing probabilistic cause-of-death assignment methods. We also develop an early stopping criterion that fully accounts for the uncertainty in the model parameters. We also propose a penalized score to account for constraints and preferences of existing question structures. We evaluate the performance of our active designs using both synthetic and real data, demonstrating that the proposed strategy achieves accurate cause-of-death assignment using considerably fewer questions than the traditional static VA survey instruments.

Data and Code Availability The data and code to replicate this paper are publicly available. We use synthetic data generated with the replication codes, and the Population Health Metrics Research Consortium (PHMRC) gold-standard VA dataset, which is publicly available at https://ghdx.healthdata.org/record/ihme-data/population-health-metrics-research-consortium-gold-standard-verbal-autopsy-data-2005-2011.

Institutional Review Board (IRB) The study does not require IRB approval.

1. Introduction

Data on cause of death is essential for understanding the heterogeneous burden of diseases. Most low- and middle-income countries, however, do not have vital statistics systems that produce high quality statistics on cause of death. As a result, about two thirds of deaths worldwide are not registered or assigned a cause (World Health Organization, 2021). Verbal autopsy (VA) is a widely used tool to collect information on cause of death when medically certified cause-of-death information is not available. VA is conducted through a structured questionnaire administered to family members or caregivers of a recently deceased person. The questionnaire collects information about the circumstances, signs, and symptoms leading up to a person's death. VAs are widely implemented in health and demographic surveillance systems, as well as national and multi-national mortality surveillance programs (see, e.g., Chandramohan et al., 2021, for an overview). Data collected by VAs can be interpreted and assigned a cause of death by physician panels, or more commonly, analyzed by statistical algorithms. Algorithmic and statistical models have been developed and routinely used to classify individual cause of death and estimate the population-level cause-specific mortality fractions. The earlier and more widely used models typically assume symptoms are conditionally independent given causes and perform cause-of-death assignment using different variations of the Naive Bayes classifier (e.g, Byass et al., 2019; McCormick et al., 2016). Several extensions to McCormick et al. (2016) have been recently proposed to further improve VA cause-of-death assignment using more complex Bayesian hierarchical models (e.g., Kunihama et al., 2020; Li et al., 2020; Moran et al., 2021). More flexible machine learning models have also been used in analyzing text-based narratives collected during VA interviews but it has been shown that they do not improve cause-of-death assignment using only the data from the structured questionnaire (Blanco et al., 2020). An overview of existing cause-of-death assignment algorithms can be found in Li et al. (2022).

Much of the literature analyzing VA data focuses on automating the process to classify causes of death. The data collection process has received much less attention in the literature. One of the main barriers for scaling up VA implementation is the challenge to conduct the overly complex and long questionnaires. For example, the current WHO 2016 standardized VA instrument includes 480 questions. While each interview evokes only a subset of the questions, a typical VA interview still needs to go through 100 to 200 questions. Lengthy interviews increase the emotional stress to both the respondents and interviewers, and can lead to survey fatigue and decreased acceptance of the interview (Loh et al., 2021; Nichols et al., 2022; Hinga et al., 2021). To the best of our knowledge, two attempts have been made in the last two decades to systematically reduce the length of VA questionnaires. Serina et al. (2015) measured the marginal associations between symptoms and causes of death, and removed symptoms based on the ranking of their importance. However, they evaluated symptom importance based on a single highly simplified classification algorithm; thus the results are heavily influenced by the parametric assumptions of the algorithm. A more recent development to simplify the VA questionnaire was carried out by the WHO in producing the 2022 standard VA instrument, described in Chandramohan et al. (2021) and Nichols et al. (2022). A more thorough process was carried out in which symptom response patterns and importance were evaluated in a model-agnostic fashion, and mixed-methods analyses were conducted to identify around 100 questions that can be removed.

Both previous approaches to shorten the VA interview are limited by the nature of the traditional survey instrument: the questionnaire needs to capture relevant symptoms associated with all potential causes of death. As the cause of death is the target of inference and unknown to the interviewer ahead of time, all questions need to be answered with the same priority by each respondent. Therefore, for a single interview, a large fraction of the collected data could potentially provide little relevant information in determining the cause of the particular death.

In this paper, we propose a statistical framework to adaptively conduct VA interviews by actively choosing the most informative questions to ask based on the collected responses, and stopping when enough information has been collected to determine the cause of death. Unlike the static screening approach to reduce the length of the questionnaire for all respondents, our approach leads to a dynamic and individualized questionnaire design that is optimized for the classification of each death. Our approach is motivated by the methods developed in the field of active learning and computerized adaptive testing. We focus our attention to its application in verbal autopsy questionnaires, as this is the first time an adaptive design is considered for the purpose of verbal autopsy questionnaires. The proposed approach, however, can also be generally useful for surveys with the aim of classifying respondents into pre-defined groups. The main contributions of our work can be summarized as follows:

1. Our active question selection strategy optimizes the cause-of-death classification of each individual death dynamically. We demonstrate that for most deaths, a small number of actively selected questions is enough to achieve the same level of classification accuracy as when using all questions.

2. Our active questionnaire design is compatible with any existing probabilistic cause-of-death assignment algorithms, and thus can be applied regardless of the choice of analysis models used to describe the joint distribution of symptoms and causes. This allows more flexibility in practice, as an analyst can choose the most appropriate analysis model to conduct the cause-of-death assignment and seamlessly adopt the proposed active questionnaire design for data collection.

3. Our approach extends existing work in psychometrics to a more principled Bayesian strategy that fully accounts for the uncertainty of the cause-of-death classification model, and we show that it leads to uncertainty-aware stopping rules that are more appropriate for high-stake tasks such as VAs.

4. We also propose a novel penalized version of the adaptive questionnaire strategy to account for practical constraints and preferences for the order of the questions.

2. Preliminaries

2.1. Active Learning

Active learning has been studied extensively in many areas of machine learning. Active learning algorithms seek to choose the optimal data instances to be used for the learning system. Most of the work in active learning focuses on choosing data points to be labeled to improve the performance of classification algorithms (Settles, 2011). Many types of data query strategies has been proposed in the literature, including uncertainty sampling (Lewis, 1995), query-by-committee (Seung et al., 1992), and approaches that aim to reduce the variance (Cohn et al., 1996) and generalization errors (Roy and McCallum, 2001). More similar to the context of this paper, active learning approaches have also been used to query complete feature vectors within a pool of observations with missing values (Melville et al., 2004; Li and Oliva, 2021). Active learning has been successfully applied to natural language modeling (Kaushal et al., 2019), computer vision (Dor et al., 2020), any many other applications. Previous work on active learning to collect survey responses is scarce. The work most related to our approach is the active matrix factorization approach for surveys measuring voter opinion proposed in Zhang et al. (2020). They developed an active question selection strategy to optimize the estimation of latent profiles of respondents under a low-rank matrix factorization model.

2.2. Computerized Adaptive Testing

Computerized adaptive testing (CAT) is a mode of testing which aims to find the optimal set of questions for each individual, thus resulting in more efficient and accurate recovery of latent traits of examinees (Weiss and Kingsbury, 1984). CAT was originally proposed for item response theory (IRT) in Lord (1971). The items are selected to maximize the test information at the current estimated ability based on IRT from an item bank. One main application of CAT is in the cognitive diagnosis models (CDMs), which is termed as cognitive diagnostic computerized adaptive testing (CD-CAT) (see, e.g., Cheng, 2009; Huebner, 2010). As CDMs have a discrete attribute space, standard CAT approaches for IRT are not directly applicable to CDMs. Several CD-CAT methods have been proposed in the literature with different item selection criteria. Two of the most widely adopted class of methods are the Shannon entropy approach, proposed by Tatsuoka (2002) and Tatsuoka and Ferguson (2003), and various procedures based on the Kullback-Leibler (KL) information (e.g. Xu et al., 2003; Cheng, 2009). Other information metrics include mutual information (Wang, 2013) and large deviation (Liu et al., 2015) are also proposed for CD-CAT. We will utilize a similar strategy based on

KL information for selecting optimal questions for VA surveys in this paper.

3. Method

3.1. Bayesian Active Questionnaire Design

We assume that there exists a question bank with J questions. Let X_{ij} denote the response to question j for death i, and $Y_i \in \{1, ..., C\}$ denote the categorical variable indicating the cause of death. We consider $X_{ij} \in \{0, 1\}$ in this paper since most of the questions collected by VA surveys are binary. The extension to general X is straightforward, and does not change the active design formulation. We consider the situation where a probabilistic model was fitted on a dataset $(X_i, Y_i)_{i=1,...,n}$ and produced estimates for the distribution $p(X, Y)$.

Our approach follows the KL information procedures in the CD-CAT literature (Cheng, 2009). In the context of VA, after t questions have been asked, we let $\mathcal{S}_t$ denote the set of questions already asked and among the questions $j \notin \mathcal{S}_t$, we identify the question with the most different distribution under the current estimated cause of death compared to alternative causes. That is, for an alternative cause y and the j-th question, we define

$$D_j(\hat{y}_i^{(t)} \parallel y) = \sum_x q_j(x \mid \hat{y}_i^{(t)}) \log \left(\frac{q_j(x \mid y_i^{(t)})}{q_j(x \mid y)} \right),$$

where $q_j(x \mid y) = p(X_{ij} = x \mid Y_i = y)$ is the conditional distribution of the j-th indicator given the cause of death being y, and $\hat{y}_i^{(t)} = \text{argmax}\, p(Y_i \mid \{X_{ij} : j \in \mathcal{S}_t\})$ is the estimated cause of death given the collected information at step t. Several different methods have been proposed to combine the KL distances to all alternative classifications in the CD-CAT literature. We adopt the idea of posterior weighted KL (PWKL) algorithm (Cheng, 2009) and maximize the weighted score for each question j defined by

$$\text{Score}_j = \sum_{y=1}^{C} D_j(\hat{y}_i^{(t)} \parallel y) p(Y_i = y \mid \{X_{ij} : j \in \mathcal{S}_t\}). \tag{1}$$

The existing CD-CAT literature typically assumes that the conditional distributions involved in computing the scores are known or can be estimated with high precision from existing data (Chang et al., 2019). In the context of VA, however, these quantities need to be estimated using a cause-of-death assignment model with usually limited training data $\mathcal{D}$. Several Bayesian methods have been introduced to infer cause of death using VAs (McCormick et al., 2016; Kunihama et al., 2020; Li et al., 2021; Wu et al., 2021) and it has been shown that considerable uncertainties exist in the classification and parameter estimations in these models. To account for the full posterior uncertainty of the conditional probabilities used to construct the PWKL score, we instead propose the following posterior predictive PWKL score

$$\text{PScore}_j = \int \text{Score}_j(\phi) p(\phi \mid \mathcal{D}) \approx \frac{1}{B} \sum_{b=1}^{B} \text{Score}_j(\phi^{(b)}),$$

where we use ϕ to denote all parameters used in the assignment model, $\phi^{(b)}$ to denote the b-th draw of ϕ from the posterior distribution $p(\phi|\mathcal{D})$ and $\text{Score}_j(\phi^{(b)})$ is the PWKL score defined in Equation 1 with $\phi^{(b)}$ plugged in. In the rest of the paper, we refer to the active question selection strategy that maximizes $\text{Score}_j(\hat{\phi})$, where $\hat{\phi}$ is the posterior mean of ϕ, as the design using point estimates and the strategy maximizing PScore_j as the design using posterior predictive scores. We note that this is a further extension that accounts for the full model uncertainty, compared to the modified PWKL score proposed in Kaplan et al. (2015), where only the uncertainty of the latent classification was integrated over.

3.2. Cause-of-Death Assignment Model

The proposed active selection strategy does not depend on any particular choice of the cause-of-death assignment model used to analyze the data, as long as the conditional probabilities in Equation 1 can be computed. In this paper, we consider a simplified version of the algorithm proposed in McCormick et al. (2016) to analyze training dataset $\mathcal{D}$. Our analysis model assumes the following data generating process

$$Y_i \sim \text{Cat}(\boldsymbol{\pi}),$$
$$p(X_i = x_i \mid Y_i = c) = \prod_j \theta_{cj}^{x_{ij}} (1 - \theta_{cj})^{1 - x_{ij}},$$

with conjugate priors $\theta_{cj} \sim \text{Be}(a_c, b_c)$ and $\boldsymbol{\pi} \sim \text{Dir}(\boldsymbol{\alpha})$ respectively. The posterior distributions of the parameters are

$$\boldsymbol{\pi} \mid \mathcal{D} \sim \text{Dir}(n_1 + \alpha_1, \ldots, n_C + \alpha_C),$$

$$\theta_{cj} \mid \mathcal{D} \sim \text{Be}\left(a_c + \sum_{i:Y_i=c} x_{ij},\ b_c + n_c - \sum_{i:Y_i=c} x_{ij}\right),$$

where $n_c = \sum_{i=1}^{n} \mathbf{1}(Y_i = c)$ for $c = 1, \ldots, C$. When some of the Y_i are unknown in the data, this assignment model can also be trivially extended to generate posterior draws of $(\boldsymbol{\pi}, \boldsymbol{\theta}, Y_{miss})$.

3.3. Stopping Criterion

In adaptive testing settings where the test duration is the main constraint, it is usually reasonable to stop the test after a pre-specified number of questions (e.g. Chen et al., 2012; Wang et al., 2011; Cheng, 2009). A fixed length stopping rule is straightforward to implement in our active VA questionnaire design as well. However, it is usually more appropriate to stop the interview only when enough precision has been achieved. Several related approaches on early stopping were developed in the literature. Tatsuoka (2002) proposed a stopping criteria where the largest probability of the classified class reached a given value, which was later adapted by Hsu et al. (2013) where another condition on the second largest probability was added. In our notation, the criterion proposed in Hsu et al. (2013) suggests stopping the questionnaire when the largest value of $p(Y_i = y \mid \{X_{ij} : j \in \mathcal{S}_t\})$ is larger than p_{1st} and the second largest value is smaller than p_{2nd} where $p_{\text{1st}} \geq p_{\text{2nd}}$ are two pre-specified thresholds.

The simplicity of this stopping rule is appealing, but when parameters are estimated with large uncertainty, stopping the questionnaire based on point estimates of classification probabilities may lead to erroneous early stopping. Instead, we compute the posterior predictive probability of meeting a pre-specified stopping rule similar to that introduced in Hsu et al. (2013). At each iteration, for the b-th posterior draw $\phi^{(b)}$, we compute $p_y^{(b)} = p(Y_i = y \mid \{X_{ij} : j \in \mathcal{S}_t\}, \phi^{(b)})$. The current most likely cause assignment is then $y^\star = \text{mode}_b(\text{argmax}_{y=1,\ldots,C}\, p_y^{(b)})$. We can compute the posterior predictive probability for the event

$$p(Y_i = y^\star \mid \{X_{ij} : j \in \mathcal{S}_t\}) > p_{\text{1st}},$$
$$p(Y_i = y \mid \{X_{ij} : j \in \mathcal{S}_t\}) < p_{\text{2nd}}, \quad \forall y \neq y^\star$$

We can then stop the survey when this probability exceeds a certain tolerance threshold $r \in (0, 1)$. More generally, while we adopt this specific stopping criteron in this paper, the fully Bayesian nature of the proposed score formulation also allows other stopping criterion to be similarly plugged in.

3.4. Accounting for Existing Flow of the Questionnaire

For traditional static VA surveys, the questionnaire structure usually follows a carefully designed order that leads to a natural flow of the questions. The dynamic nature of the active questionnaire design inevitably breaks such an ordering of questions, and may lead to consecutive questions that are concerned with very different aspects of an individual's death. While this is desired from the perspective of maximizing the collected information quickly, one practical concern for jumping across different topics is that it may increase the chance of inaccurate responses from the respondents. It is straightforward to impose deterministic skip patterns by modifying the search space $\mathcal{S}_t$ at each iteration based on the collected responses. For example, VA surveys typical include questions that are only triggered when a root question has been answered. We may let $\mathcal{S}_0$ include only root-questions and any sub-questions are added to $\mathcal{S}_t$ only after the corresponding root-question has been answered. In addition, it may also be useful to maintain some of the natural ordering of the questions in a 'soft' fashion. The proposed active design strategy can be easily extended to incorporate such preference by adding a penalization for certain moves across the questions, i.e., let $j^{(t)}$ denote the index of question asked in the t-th iteration, and define the penalized score

$$\text{PScore}'_j = \text{PScore}_j - \lambda D(j, j^{(t)})$$

where $\lambda > 0$ is a parameter that regulates the degree of penalization for the jumping behavior and $D(j, k)$ is a pre-specified distance metric between the j-th and k-th questions. For example, when there is a group of questions that we would like to ask together but have no preference for the order within the group, we may let $D(j,k) = 0$ if the questions are within the same group and 1 otherwise.

4. Experiments

4.1. Synthetic Data

We first generate synthetic data to evaluate the performance of the active questionnaire design strategy. We consider the following two data generating processes:

1. Correctly specified model: we generate observations using the model described in Section 3.2. We let $C = 10$, $J = 50$, $\boldsymbol{\alpha} = (1, ..., 1)$ and (a_c, b_c) to be $(0.5, 0.5)$ for $c = 1, 2, 3$, $(3, 3)$ for $c = 4, 5, 6$, and $(1, 3)$ for $c = 7, 8, 9, 10$.

2. Misspecified model: we generate observations by a latent class model such that each cause of death consist of multiple unobserved sub-categories. That is, we generate $Z_i \mid Y_i = c \sim \text{Cat}(\boldsymbol{\lambda}_c)$ and

$$p(X_i = \boldsymbol{x}_i | Y_i = c, Z_i = k) = \prod_j \theta_{ckj}^{x_{ij}} (1-\theta_{ckj})^{1-x_{ij}}.$$

   We let $\boldsymbol{\lambda}_c = (\lambda_{c1}, \lambda_{c2}, \lambda_{c3}) \sim \text{Dir}(1, 1, 1)$ and $\theta_{ckj} \sim \text{Beta}(1, 1)$ for all c, k, and j.

In each case, we generate $n_1 = 200$ and 1000 training observations, respectively, and evaluate the performance of different questionnaire designs on $n_0 = 200$ test observations.

In the first experiment, we consider running the questionnaire with a fixed number of questions. Figure 1 shows the probability of correct classification of cause of death in the test set, given different lengths of the questionnaire. In all cases, the two active question selection strategies reaches high classification accuracy faster than asking the questions sequentially with a fixed order. In the case of the correctly specified model, we also evaluate the accuracy of the oracle strategy when the parameters of the true data generating process are known, and both active strategies perform similarly to the oracle when the training data is sufficiently large.

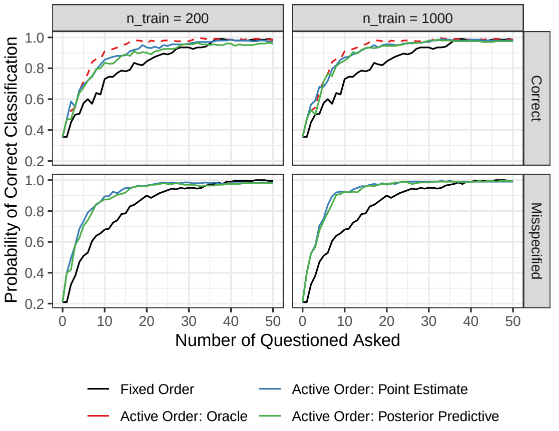

Figure 1: Classification accuracy of different questionnaire design as the the number of questions increases.

We then evaluate the performance of the active design under different varying-length stopping rules under the same data generating processes as in the first experiment. We compare the classification accuracy of different active design strategies under the two data generating processes, respectively, when different stopping criteria are satisfied. We consider three stopping rules: when using the point estimates to select questions, we consider the same criterion in Hsu et al. (2013); when using the posterior predictive scores, we consider stopping when $r = 50\%$ and 70% of the posterior draws meet the same criterion. Table 1 and 2 summarize the results. We fix $p_{1\text{st}} = 0.8$ and 0.9, and for the second threshold, we compute $p_{2\text{nd}} = (1 - p_{1\text{st}})/C + d(C-2)(1-p_{1\text{st}})/C$ for various choices of $d \in [0, 1]$, as suggested by Hsu et al. (2013). Overall, classification accuracy increases as we increase $p_{1\text{st}}$ and decrease $p_{2\text{nd}}$ (or equivalently decrease d), as more questions need to be asked before the more stringent stopping criterion are met. Due to the space limitation, we present results for $d = 0$ and 0.5 only. In addition to accuracy, we also compare the median, and 5th and 95th percentiles of the questionnaire length. It is worth noting that in Table 2, when the analysis model is misspecified, the stopping rule with 70% posterior probability of satisfying the stopping criterion leads to similar median questionnaire length compared to considering only point estimates, while also having a longer tail in the number of questions asked. This leads to higher overall accuracy. In high-stake contexts such as VA, it is often more appropriate to favor conservative strategies and collect more information for deaths that are difficult to classify, especially when the analysis model driving the questionnaire design is not accurate enough. In such cases, the proposed stopping rule with a larger r might be preferred in practice.

Table 1: Classification accuracy and questionnaire length for the synthetic data under the correctly specified model. Median, 5th and 95th percentile of the questionnaire length are shown for each stopping rule conditions.

p_{1st}	d	Stopping Rule	Acc	Median	Lower	Upper
0.8	0.5	Point Est	0.85	5	3	14
		Pred $r = 0.5$	0.92	10	4	50
		Pred $r = 0.7$	0.95	14	6	50
0.8	0	Point Est	0.96	8	5	35
		Pred $r = 0.5$	0.96	13	6	50
		Pred $r = 0.7$	0.96	20	7	50
0.9	0.5	Point Est	0.95	7	5	18
		Pred $r = 0.5$	0.95	12	6	50
		Pred $r = 0.7$	0.96	18	7	50
0.9	0	Point Est	0.97	10	7	50
		Pred $r = 0.5$	0.96	16	8	50
		Pred $r = 0.7$	0.96	23	8	50

Finally, we consider situations where the responses may be subject to errors when the questionnaire deviates from a natural flow. Let $j^{(t)}$ denote the index for the t-th question asked, we simulate noisy responses during the interview by letting

$$X^{\star}_{ij^{(t)}} = \begin{cases} 1 - X_{ij^{(t)}} & \text{w.p. } \frac{D(j^{(t-1)}, j^{(t)})}{h}, \\ X_{ij^{(t)}} & \text{w.p. } 1 - \frac{D(j^{(t-1)}, j^{(t)})}{h} \end{cases}$$

Here we consider the distance metric $D(j, k) = |j - k|/p$. In this setup, asking questions sequentially leads to the least amount of added noise. We compare different active questionnaire designs using the penalized score described in Section 3.4 with $\lambda = 2$ and 10. We consider the same data generating process as in the first experiment, and both a low noise setting ($h = 10$) and a high noise setting ($h = 2$). Figure 2 shows that the unpenalized active designs lead to suboptimal classification accuracy in high noise settings, as the responses include more errors induced by the non-sequential order of the questions. Active designs based on the proposed penalized score are able to mitigate the effect of response errors and achieve classification accuracy comparable to or higher than static questionnaire designs with the optimal order.

Table 2: Classification accuracy and questionnaire length for the synthetic data under the misspecified model. Median, 5th and 95th percentile of the questionnaire length are shown for each stopping rule conditions.

p_{1st}	d	Stopping Rule	Acc	Median	Lower	Upper
0.8	0.5	Point Est	0.88	5	3	11
		Pred $r = 0.5$	0.88	4	3	18
		Pred $r = 0.7$	0.94	5	3	27
0.8	0	Point Est	0.96	6	5	17
		Pred $r = 0.5$	0.94	6	4	30
		Pred $r = 0.7$	0.99	8	5	50
0.9	0.5	Point Est	0.94	6	4	13
		Pred $r = 0.5$	0.96	6	4	23
		Pred $r = 0.7$	0.98	7	4	50
0.9	0	Point Est	0.96	7	6	20
		Pred $r = 0.5$	0.98	8	5	42
		Pred $r = 0.7$	0.99	9	5	50

4.2. PHMRC Data

In this section, we consider the application of the adaptive questionnaire strategy on the Population Health Metrics Research Consortium (PHMRC) gold-standard VA dataset (Murray et al., 2011). The PHMRC dataset is widely used for validating VA cause-of-death assignment methods (McCormick et al., 2016; Kunihama et al., 2020; Moran et al., 2021). It consists of 7,841 adult deaths collected from six study sites (Andhra Pradesh, India; Bohol,

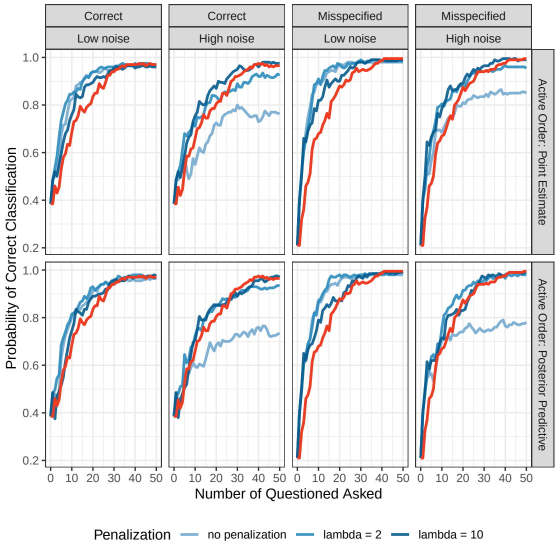

Figure 2: Classification accuracy of different questionnaire design when responses include order-induced noises. The red curves correspond to fixed sequential order.

Philippines; Dar es Salaam, Tanzania; Mexico City, Mexico; Pemba Island, Tanzania; and Uttar Pradesh, India). Gold-standard causes were determined based on laboratory, pathology and medical imaging findings. In this dataset, there are $C = 34$ cause-of-death categories and we preprocessed the raw dataset into $J = 168$ binary indicators using the steps described in Li et al. (2022).

We again consider both stopping at fixed length and when using the varying-length stopping rules. We conduct a 10-fold cross validation analysis, where we evaluate different active questionnaire strategies on each fold of data with the rest of data being used to estimate model parameters. We treat missing values in the dataset to be missing at random when fitting the model, as is commonly assumed in existing VA cause-of-death assignment algorithms (McCormick et al., 2016; Kunihama et al., 2020; Li et al., 2021).

Figure 3 shows the accuracy of the active questionnaire using point estimates and posterior predictive scores, and compare them with the traditional questionnaire with fixed question order. Both strategies achieve a considerably higher classification accuracy compared to the static design after as few as 10 questions. In fact, the overall accuracy of the active designs is optimal when around 30 to 40 questions have been asked, and then the accuracy slightly decreases as the questionnaire becomes longer. This is not surprising, as the analysis model in this experiment is likely overly simplistic and does not approximate the complex distribution of symptoms and causes well in the real data. However, while the design of better cause-of-death assignment algorithms remains an important research topic, the experiment clearly demonstrates that even with a simple analysis model, the active questionnaire design can lead to highly accurate cause-of-death classifications using only 1/4 of the questions.

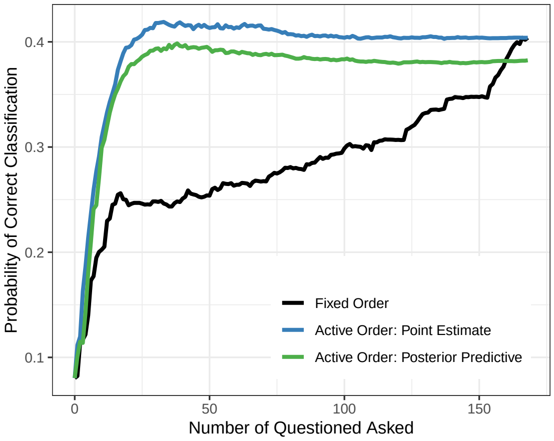

Figure 3: Classification accuracy of different questionnaire designs on the PHMRC data as the the number of questions increases.

Lastly, we apply the adaptive early-stopping rule to the same cross validation experiment. We consider $p_{1\text{st}} = 0.7$, 0.8, and 0.9 and fix $d = 0.5$. We present additional sensitivity analysis using different values of d in the supplementary materials. We again compare the stopping rules based on the point estimate, and posterior predictive probabilities with $r = 0.5$ and 0.7. Figure 4 examines the relationship between the proportion of correctly specified deaths when the questionnaire stops and the median number of questions

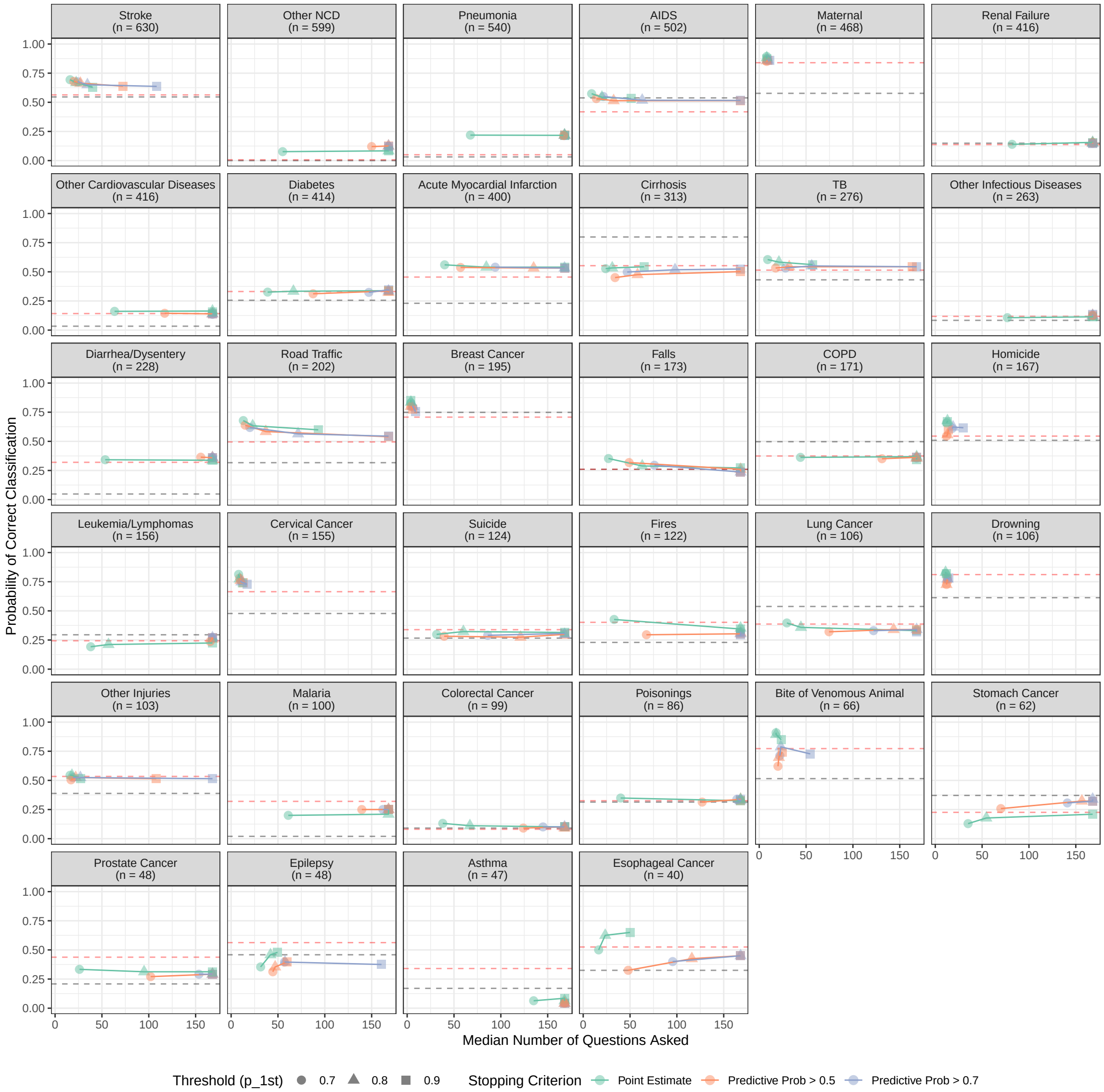

Figure 4: Proportion of correctly classified deaths among deaths due to each cause using different stopping criterion. The causes are ordered by their sample size in the PHMRC data. The dotted horizontal lines correspond to accuracy from two widely adopted VA cause-of-death assignment algorithms using all symptoms. Red line: InSilicoVA (McCormick et al., 2016). Blue line: InterVA (Byass et al., 2019).

asked, disaggregated by the true cause of death. For most causes, classification accuracy is not sensitive to the choice of thresholds we use and only the length of the conducted questionnaire changes as the thresholds become more stringent. This is as expected from the observations in the first experiment that only a small number of questions is needed to achieve high classification accuracy. However, we note that for causes such as maternal deaths, breast cancer, cervical cancer, drowning, homicide, and bite of venomous animals, the active questionnaire design achieves high classification accuracy quickly regardless of stopping criterion. This is likely due to the strong associations between these causes and a small number of key symptoms. This observation is highly useful in practice as it allows interviewers to terminate the VA interview considerably sooner when sufficient information has been collected to identify certain causes that are easier to classify. We also compare the performance of the active questionnaire strategy with two widely adopted VA cause-of-death assignment methods, InSilicoVA (McCormick et al., 2016) and InterVA (Byass et al., 2019) in Figure 4. The active questionnaire strategy is able to achieve comparable and usually higher accuracy compared to both of the state-of-the-art VA algorithms using the full dataset.

5. Discussion

In this paper, we introduced a novel active questionnaire design strategy for verbal autopsy surveys. We proposed a principled Bayesian formulation to estimate posterior predictive scores of questions based on the KL information of questions in the bank. Our approach takes into account the uncertainty from any probabilistic cause-of-death assignment models and can facilitate adaptive early stopping rules and incorporate the existing flow of the questionnaire. We demonstrated improved performance on cause-of-death classification with both fixed and adaptive length of the questionnaire. More broadly, while we focus on verbal autopsy surveys in this work, the same methodology can be applied to other medical and health surveys in resources-constrained settings.

The proposed active questionnaire can be readily adopted in the field as electronic data collection through tablets is already the standard practice for VA field interviews (World Health Organization, 2022). The computation of the analysis model only needs to be performed once, or updated routinely, in a separate backend before data collection. The only computation required for real-time question selection involves the computation of Equation 1 given pre-computed parameter values, which can be efficiently implemented on existing tablets. The choice of the analysis model is contextual and depends on the implemented pipeline of VA data analysis. We anticipate that field experiments are needed to determine the choice of tuning parameters for early termination of the questionnaires, which is beyond the scope of this paper.

While this work provides a novel approach to questionnaire design in VA studies, we acknowledge several limitations for the approach to be directly useful in practice. First, the analysis model we used in this paper is over simplified compared to the more recent work in the literature (e.g., Kunihama et al., 2020; Li et al., 2021; Wu et al., 2021). As a result, while we observe advantage in the posterior predictive scores in determining early stopping rules, the overall classification accuracy is not significantly different from active ordering based on only the point estimates of the parameters. Combining the active design with more complex analysis models is beyond the scope of this work and is an important future direction. Second, there is extensive domain knowledge on the relationship and logic behind the questions on VA surveys. Such information may allow researchers to construct more useful penalty functions to regulate the flow of the active questionnaire and provide guidance on choosing the tuning parameter for the penalty.

Several future directions of research could further address the methodological challenges of active questionnaire design for VAs. First, this work focuses on the situation where a cause-of-death assignment model has been chosen to analyze existing data. The active questionnaire strategy may be further improved to account for more than one analysis models. Ensemble prediction has been shown to improve the performance of cause-of-death classification (Datta et al., 2021) and could lead to active questionnaire designs and stopping rules that are more robust to model misspecification. Second, the recent work on domain adaptive VA algorithms illustrates that VA data are typically heterogeneous across different populations and the model parameters fitted with data from one population may not lead to good predictions for another population (Li et al., 2021; Wu et al., 2021). The proposed adaptive design could be adapted to also account for this additional layer of uncertainty. Third, while we focus on the task of optimizing questionnaire design for future data collection given models estimated on existing data, it remains an open question how to efficiently combine and jointly analyze data collected via different adaptive and traditional instruments. Finally, it is an important practical research area to understand how VA interviewees respond to questionnaires with different orders and the potential impact of instrument on data quality. Furthermore, VA questions have different emotional burdens on the respondents, take varying amounts of time to conduct, and are subject to different types of bias across populations. It is also important to quantify the differential cost of each question to better understand the trade-off between classification accuracy and cost of the interview. We leave these directions for future research.

References

Alberto Blanco, Alicia Perez, Arantza Casillas, and Daniel Cobos. Extracting cause of death from verbal autopsy with deep learning interpretable methods. *IEEE Journal of Biomedical and Health Informatics*, 2020.

Peter Byass, Laith Hussain-Alkhateeb, Lucia D'Ambruoso, Samuel Clark, Justine Davies, Edward Fottrell, Jon Bird, Chodziwadziwa Kabudula, Stephen Tollman, Kathleen Kahn, Schiőler Linus, and Max Petzold. An integrated approach to processing WHO-2016 verbal autopsy data: the InterVA-5 model. *BMC Medicine*, 17(1):1–12, 2019.

Daniel Chandramohan, Edward Fottrell, Jordana Leitao, Erin Nichols, Samuel J. Clark, Carine Alsokhn, Daniel Cobos Munoz, Carla AbouZahr, Aurelio Di Pasquale, Robert Mswia, Eungang Choi, Frank Baiden, Jason Thomas, Isaac Lyatuu, Zehang Richard Li, Patrick Larbi-Debrah, Yue Chu, Samuel Cheburet, Osman Sankoh, Azza Mohamed Badr, Doris Ma Fat, Philip Setel, Robert Jakob, and Don de Savigny. Estimating causes of death where there is no medical certification: evolution and state of the art of verbal autopsy. *Global Health Action*, 14(sup1):1982486, 2021. doi: 10.1080/16549716.2021.1982486.

Yuan-Pei Chang, Chia-Yi Chiu, and Rung-Ching Tsai. Nonparametric CAT for CD in educational settings with small samples. *Applied Psychological Measurement*, 43(7):543–561, 2019.

Ping Chen, Tao Xin, Chun Wang, and Hua-Hua Chang. Online calibration methods for the DINA model with independent attributes in CD-CAT. *Psychometrika*, 77(2):201–222, April 2012. ISSN 1860-0980. doi: 10.1007/s11336-012-9255-7.

Ying Cheng. When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika*, 74(4):619, April 2009. ISSN 1860-0980. doi: 10.1007/s11336-009-9123-2.

David A Cohn, Zoubin Ghahramani, and Michael I Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145, 1996.

Abhirup Datta, Jacob Fiksel, Agbessi Amouzou, and Scott L Zeger. Regularized Bayesian transfer learning for population-level etiological distributions. *Biostatistics*, 22(4):836–857, October 2021. ISSN 1465-4644. doi: 10.1093/biostatistics/kxaa001.

Liat Ein Dor, Alon Halfon, Ariel Gera, Eyal Shnarch, Lena Dankin, Leshem Choshen, Marina Danilevsky, Ranit Aharonov, Yoav Katz, and Noam Slonim. Active learning for BERT: An empirical study. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7949–7962, 2020.

Alex Hinga, Vicki Marsh, Amek Nyaguara, Marylene Wamukoya, and Sassy Molyneux. The ethical implications of verbal autopsy: responding to emotional and moral distress. *BMC Medical Ethics*, 22(1):1–16, 2021.

Chia-Ling Hsu, Wen-Chung Wang, and Shu-Ying Chen. Variable-Length Computerized Adaptive Testing Based on Cognitive Diagnosis Models. *Applied Psychological Measurement*, 37(7):563–582, October 2013. ISSN 0146-6216. doi: 10.1177/0146621613488642.

Alan Huebner. An overview of recent developments in cognitive diagnostic computer adaptive assessments. *Practical Assessment, Research, and Evaluation*, 15(1):3, 2010.

Mehmet Kaplan, Jimmy de la Torre, and Juan Barrada. New item selection methods for cognitive diagnosis computerized adaptive testing. *Applied Psychological Measurement*, 39 (3):167–188, 2015. ISSN 0146-6216. doi: 10.1177/0146621614554650.

Vishal Kaushal, Rishabh Iyer, Suraj Kothawade, Rohan Mahadev, Khoshrav Doctor, and Ganesh Ramakrishnan. Learning from less data: A unified data subset selection and active learning framework for computer vision. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1289–1299. IEEE, 2019.

Tsuyoshi Kunihama, Zehang Richard Li, Samuel J. Clark, and Tyler H. McCormick. Bayesian factor models for probabilistic cause of death assessment with verbal autopsies. *The Annals of Applied Statistics*, 14(1):241–256, March 2020. doi: 10.1214/19-AOAS1253.

David D Lewis. A sequential algorithm for training text classifiers: Corrigendum and additional data. In *Acm Sigir Forum*, volume 29, pages 13–19. ACM New York, NY, USA, 1995.

Yang Li and Junier Oliva. Active feature acquisition with generative surrogate models. In *International Conference on Machine Learning*, pages 6450–6459. PMLR, 2021.

Zehang R Li, Tyler H McCormick, and Samuel J Clark. Using Bayesian latent Gaussian graphical models to infer symptom associations in verbal autopsies. *Bayesian Analysis*, 15(3):781, 2020.

Zehang Richard Li, Zhenke Wu, Irena Chen, and Samuel J Clark. Bayesian nested latent class models for cause-of-death assignment using verbal autopsies across multiple domains. *arXiv preprint arXiv:2112.12186*, 2021.

Zehang Richard Li, Jason Thomas, Eungang Choi, Tyler H McCormick, and Samuel J Clark. The openva toolkit for verbal autopsies. *arXiv preprint arXiv:2109.0824*, 2022.

Jingchen Liu, Zhiliang Ying, and Stephanie Zhang. A rate function approach to computerized adaptive testing for cognitive diagnosis. *Psychometrika*, 80(2):468–490, 2015.

Patricia Loh, Edward Fottrell, James Beard, Naor Bar-Zeev, Tambosi Phiri, Masford Banda, Charles Makwenda, Jon Bird, and Carina King. Added value of an open narrative in verbal autopsies: a mixed-methods evaluation from malawi. *BMJ Paediatrics Open*, 5 (1), 2021.

Frederic M Lord. Robbins-monro procedures for tailored testing. *Educational and Psychological Measurement*, 31(1):3–31, 1971.

Tyler H. McCormick, Zehang Richard Li, Clara Calvert, Amelia C. Crampin, Kathleen Kahn, and Samuel J. Clark. Probabilistic Cause-of-Death Assignment Using Verbal Autopsies. *Journal of the American Statistical Association*, 111(515):1036–1049, July 2016. ISSN 0162-1459. doi: 10.1080/01621459.2016.1152 191.

Prem Melville, Maytal Saar-Tsechansky, Foster Provost, and Raymond Mooney. Active feature-value acquisition for classifier induction. In *Fourth IEEE International Conference on Data Mining (ICDM'04)*, pages 483–486. IEEE, 2004.

Kelly R. Moran, Elizabeth L. Turner, David Dunson, and Amy H. Herring. Bayesian hierarchical factor regression models to infer cause of death from verbal autopsy data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 70(3):532–557, June 2021. ISSN 0035-9254. doi: 10.1111/rssc.12468.

Christopher JL Murray, Alan D Lopez, Robert Black, Ramesh Ahuja, Said M Ali, Abdullah Baqui, Lalit Dandona, Emily Dantzer, Vinita Das, Usha Dhingra, et al. Population health metrics research consortium gold standard verbal autopsy validation study: design, implementation, and development of analysis datasets. *Population health metrics*, 9(1):27, 2011.

Erin Nichols, Kristen Pettrone, Brent Vickers, Hermon Gebrehiwet, Clarissa Surek-Clark, Jordana Leitao, Agbessi Amouzou, Dianna M Blau, Debbie Bradshaw, El Marnissi Abdelilah, Pamela Groenewald, Brian Munkombwe, Chomba Mwango, F. Sam Notzon, Steve Biko Odhiambo, and Paul Scanlon. Mixed-methods analysis of select issues reported in the 2016 world health organization verbal autopsy questionnaire. *Plos one*, 17(10): e0274304, 2022.

Nicholas Roy and Andrew McCallum. Toward optimal active learning through monte carlo estimation of error reduction. *ICML, Williamstown*, 2:441–448, 2001.

Peter Serina, Ian Riley, Andrea Stewart, Abraham D Flaxman, Rafael Lozano, Meghan D Mooney, Richard Luning, Bernardo Hernandez, Robert Black, Ramesh Ahuja, et al. A shortened verbal autopsy instrument for use in routine mortality surveillance systems. *BMC Medicine*, 13:1–10, 2015.

Burr Settles. From theories to queries: Active learning in practice. In *Active learning and experimental design workshop in conjunction with AISTATS 2010*, pages 1–18. JMLR Workshop and Conference Proceedings, 2011.

H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 287–294, 1992.

Curtis Tatsuoka. Data analytic methods for latent partially ordered classification models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 51(3):337–350, February 2002. ISSN 0035-9254. doi: 10.1111/1467-9876.00272.

Curtis Tatsuoka and Thomas Ferguson. Sequential classification on partially ordered sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):143–157, February 2003. ISSN 1369-7412. doi: 10.1111/1467-9868.00377.

Chun Wang. Mutual information item selection method in cognitive diagnostic computerized adaptive testing with short test length. *Educational and Psychological Measurement*, 73 (6):1017–1035, 2013.

Chun Wang, Hua-Hua Chang, and Alan Huebner. Restrictive stochastic item selection methods in cognitive diagnostic computerized adaptive testing. *Journal of Educational Measurement*, 48(3):255–273, September 2011. ISSN 0022-0655. doi: 10.1111/j.1745-3984.2011.00145.x.

David J Weiss and G Gage Kingsbury. Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, 21(4):361–375, 1984.

World Health Organization. WHO civil registration and vital statistics strategic implementation plan 2021-2025, 2021.

World Health Organization. Verbal autopsy standards: verbal autopsy field interviewer manual for the 2022 WHO verbal autopsy instrument, 2022.

Zhenke Wu, Zehang Richard Li, Irena Chen, and Mengbing Li. Tree-informed Bayesian multi-source domain adaptation: cross-population probabilistic cause-of-death assignment using verbal autopsy. *arXiv preprint arXiv:2112.10978*, 2021.

Xueli Xu, H Chang, and Jeff Douglas. A simulation study to compare CAT strategies for cognitive diagnosis. In *annual meeting of the American Educational Research Association, Chicago*, page 34, 2003.

Chelsea Zhang, Sean J. Taylor, Curtiss Cobb, and Jasjeet Sekhon. Active matrix factorization for surveys. *The Annals of Applied Statistics*, 14(3):1182 – 1206, 2020. doi: 10.1214/20-AOAS1322.